Turkish Journal of Computer and Mathematics Education *Vol.12 No.6 (2021), 4219-4225*

Research Article

Fake News Detection in low-resourced languages "Kurdish language" using Machine learning algorithms

Rania Azad[a,b] , Bilal Mohammed[a], Rawaz Mahmud[a], Lanya Zrar[a], Shajwan Sdiq[a]
[a]Network department, College of informatics, Sulaimani Polytechnic University, Sulaimani, Iraq[1]
[b]American University of Iraq, Sulaimani, Iraq[2]

Article History: *Do not touch during review process(xxxx)*

Abstract: With the growth of using the internet and the large amount of real-time information created and shared over social media platforms, the risk of disseminating malicious activities, perform illegal movements, abuse other people, and publicize fake news increased dramatically. Fake news detection is a well-studied research issue to understand the nature of fake news, detection or prevention for the highly resourced languages like Arabic, English, and other European languages where less-resourced languages remain out of the focus because of the absence of labeled fake corpus, absence of fact-checker websites or unavailability of NPL tools, until today, non-research has been conducted in Fake news detection in the Kurdish language. This paper showcase creating a novel Kurdish Fake news corpus that made publicly available[1],it contains two sets of news, the first one contains crawled fake news, the second set contains manipulated text from real news, then several classifiers applied on the corpus after using TF-IDF as a feature of selection. The outcome of the proposed paper showed that Support Vector Machine (SVM) scored the highest accuracy 88.71% among the other classifiers on set 1 and LR outperforms the other algorithms on set 2. This work can be considered as a baseline for future studies.

Keywords: Fake News Detection, Kurdish Language, Machine learning, Classifiers, SVM, TF-IDF.

1. Introduction

Over the past years, the use of social media has greatly increased, with its benefits of connecting people, sharing content, and staying connected about worldwide events(Zhou, Cai, Zeng, & Wang, 2020), the risk of misleading information and the danger of spreading fake news (FN) increased which potentially might cause serious problems in society(Zhang & Ghorbani, 2020)Several definitions exist in the literature, andmost of them agreed that Fake news refers to fabricated information intentionally tomislead, befool or lure readers for financial, political, or other gains purposely[1-3]. It is considered that the area of fake news detection is one of the most difficult and sensitive issues in Natural Language Processing (NLP) due to the variety of news sources, the language used, and its pattern. Additionally, the variety of text transformation techniques types used, and machine learning algorithms applied to address the problem.After U.S. presidential 2016, Fake news detection systems become an emerging topic in the research area (B. Collins, D. T. Hoang, N. T. Nguyen, 2020). Many companies were seeking future solutions for identifying fake information such as Facebook, Google(Zhang & Ghorbani, 2020).

On the other hand,The majority of the previous studies got focused on the English language due to the availability of well-known annotated fake corpus openly available, variety of fact-checkers around the world while the less-resourcedlanguagesleft behind such asthe Kurdish language.While the Kurdish language is spoken by more than 30 million people around the world (Abdulrahman, Hassani, & Ahmadi, 2019), yet, it is considered as less-resourced in the Natural Language Processing (NLP) domain due to the inaccessibility of NLP tools and theshortage or unavailability of the labeled corpus.(Sharma, Litoriya, Pratap Singh, & Sharma, 2021)
It is worth mentioning that the Kurdish language is an Indo-European language mainly spoken in central and eastern Turkey, Western Iran, northern Iraq, and Syria. Sorani dialect has been chosen in this study as it is among the main five Kurdish dialects spoken in our region Iraq. [4-5-6]

The key contributions of this research are summarized as follows:

- To the best of our knowledge, it might be the first study that focuses on fake news detection written in the Kurdish language.

[1]https://www.kaggle.com/raniaazad/kurdish-fake-news-dataset

- A Kurdish fake news corpus developed and made publicly available for further research which contains two different sets(crawled fake news from illegitimate sources and manipulated text).
- Applied Five machine learning models on the corpus after using TD-IDF vectorizer as feature selection.

This paper's structure is set as follows: in section 2, we staterelated-work and what has been done in a similar issue, the methodology of the experiment is described in section 3, followed by section 4 that discusses the result and analysis of fiveclassifiers then we conclude the paper with section 6.

2.Related work

In this section, an extensive investigation is presented in an attempt to understand the achieved works in identifying FN on less-resourced languages that itswriting orientation is from right to left like Arabic, and Urdu,which are considered as similar forms and orientation to the Kurdish language.[4-5]

Authors(Maysoon Alkhair, Karima Meftouh & Othman, 2019)investigated three classifiers methods: Decision Tree (DT), Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) on Arabic content crawled from YouTube comments using specific keywords. an n-gram of words has been used as features endorsed by TF-IDF, the best results in terms of the best recall are achieved by the decision tree and the accuracy and precision are obtained by the SVM.
Furthermore, (Alanazi, 2020)proposed credibility verification by using text mining techniques on Arabic Facebook comments collected to create a dataset, as a result, the Naïve Bayes (NB) classifier achieved an accuracy of 87.18% against the Support Vector Machine (SVM) classifier 87.14%.while (AL-Saif & Al-Dossari, 2018) proposed detecting Arabic crimes by using different classification algorithms: SVM, DT, CNB, and KNN, The experiment result revealed that SVM with tri-gram scored the highest accuracy 91.55%.A recent study case by (Maakoul et al., 2020) applied Logistic Regression (LR) classifier oncrawled comments from one specific Facebookfake post, the accuracy was 62%.More recent work done by (Hybrid, Hawks, & Feature, 2021)proposedintelligent detection of fake informationon Arabic Twitter corpus using Natural Language Processing (NLP) techniques, 8 well-known algorithms, and Harris Hawks Optimizer (HHO) as a wrapper-based feature selection approach. Thefindingspresented that the Logistic Regression (LR) with Term Frequency-Inverse Document Frequency (TF-IDF) model scores the best rank.Another approach has been used by (Elmurngi & Gherbi, 2017), analyzed a movie reviews corpus using Sentiment Analysis and text classification methods, then compare five supervised machine learning algorithms: Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN-IBK), KStar (K*), and Decision Tree (DT-J48), SVM algorithm outperforms other algorithms by 76%.

An advanced work has been pursued by (Nagoudi, Elmadany, Abdul-Mageed, Alhindi, & Cavusoglu, 2020), proposed a novel approach to detect machine generation and manipulated text on different Arabic datasets with POS tagging, the results showed an F1 = 70,06% using 4 masks: mBERT, XLM-RBase, XLM-RLarge, and AraBERT.Anotherenhanced work done by (Khalifa & Hussein, 2019) used classical, deep, and hybrid ensembleswith 5 extracted featuresinclude TF-IDF word n-gram features, bag-of-words representation, sentiment-based features, and topic modeling features. The findings show the classical ensemblebeating both deep and hybrid ensembles with 84.4 F1 points

While in the Urdu language, authors reached similar contributions like (Amjad et al., 2020) that provided a manually compiled and validated dataset containing 900 news articles with the best accuracy 87% using AdaBoost classifier against NB, LR, RF, DT. And (Hussain, Rashidul Hasan, Rahman, Protim, & Al Hasan, 2020)reported that SVM with linear kernel scores an accuracy of 96.64% higher than MNB with a 93.32%. similar work done by (Ahmed, Traore, & Saad, 2017) that comparestwo different feature extraction techniques and six different machine classificationtechniques. The outcome of the experimental evaluation yields that the best performance is by usingTF-IDF as feature extractiontechnique, and Linear Support Vector Machine (LSVM) as a classifier, with anaccuracy of 92%.

3.Methodology

In this section, the proposed methodology to address the problem of detecting fake news and manipulated news in the Kurdish languageis summarized in Figure 1

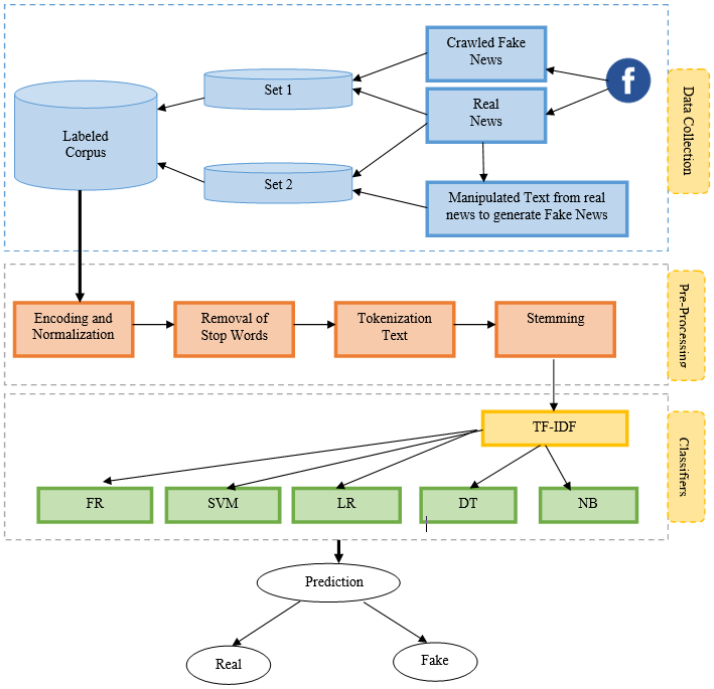

Figure 1: The proposed framework for Kurdish Fake News Detection

3.1.Data Collection

An important component of the research, is data collection, due to the absence of fake News corpus in the Kurdish language, it was mandatory to create oneentitled "KurdFake", The data is obtained through two steps:

Step 1:

5000 real news was harvested from 4 reliable and official Facebook pages in the Kurdistan region, Iraq: Rudaw[2], K24[3], NRT[4], KNN[5] using a facebook-scraper tool from different categories: Finance, Economy, Health, politics, Culture Sports, and Technology.

Step 2:

Collecting fake news was a challenging task as it required a tremendous amount of work and time to evaluate and annotate the news. Moreover, there are no fact-checking websites in the Kurdish language, and we tried to avoid translated fake news from previous datasets, moreover, collecting data from Facebook is not an easy task due to restrictions to obtain API authorization.To save time, two techniques were used to create two sets of data:

Set one: 5000 news were crawled from non-legitimate Facebook pages that match the below conditions to evaluate fake pages:

- Containing unknown link or unrealistic photo
- The title of the post doesn't match the content
- Non-legitimate sources

Set two: 5000 news automatically manipulated and modified from real news using a python script that substitutes one or more words from each news to create a contradictory meaning.

[2]https://www.rudaw.net/
[3]https://www.kurdistan24.net/
[4]https://nrttv.com/
[5]https://www.knnc.net/

Later, each set was merged with the same real news to obtain 10 000 annotated news for set one and 10 000 annotated news for set two.

3.2.Pre-processing:

Processing is a crucial step in NLP to improve the quality of text data before feeding the classifiers by removing irrelevant data before feature extraction.In this study, Kurdish Language Processing Toolkit KLPT(Ahmadi, 2020) was used for this purpose:

Data cleaning, UTF-8 encoding, and normalization

In order to improve the quality of text data and ensure the reliability of the statistical analysis tohave a relevant analysis, it is essential to clean data from special characters such as: *{*,@,%,&...}*, *URLs*, words in foreign languages, emojis, extra spaces. then convert the text to UTF-8 Unicode.

Tokenization

Tokenization is a mandatory technique since textual documents in natural language are
usually composed of long, complicated, and malformed sentences, so the aim is to split the news into a series ofsingle words separated by white space.

Removal of Stop Words

Eliminate meaningless wordssuch as connectors, articles, andpronounsusing a list of Kurdishstopwords that contains 240 stop words. (Mustafa & Rashid, 2018)

Stemming

Process returning a word to its root form or origin to improve the performance of text extraction.

3.3.Feature Extraction

Feature extraction is a significant stage to select the appropriate feature sets.Term Frequency-Inverse Document Frequency (TF-IDF)has been used commonly in literatureto transform the text into numerical values which can be fed to a machinelearning model for processing. itprovides insights about the less relevant and more relevant words in a document. It is considered a simple technique

3.4.Experiment Setup

In this paper, based on the related work section (Maakoul et al., 2020)(Maysoon Alkhair, Karima Meftouh & Othman, 2019)(Alanazi, 2020)(AL-Saif & Al-Dossari, 2018)(Elmurngi & Gherbi, 2017)(Amjad et al., 2020)(Hussain et al., 2020)and (Hussain et al., 2020), we chooseto examine the performance of five classifiers: Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) using TF-IDF feature extraction.

The experiment was implemented under the same computing environment and coded in Python that facilitates the implementation of NLP, Machine Learning, using different libraries including KLPT, Pandas, BeautifulSoup, and SKlearn.

It is worth mentioning that, each set of the corpus was split into train and test sets using 70% for train and the rest for test.

3.5.Metrics

Commonly, the evaluation of classifiers is done using different metrics based on the confusion matrix which is a tabularrepresentation of a classification model performance on the test set, which consists of four parameters: true positive, false positive, true negative, and false negative(Ahmad, Yousaf, Yousaf, & Ahmad, 2020)(Hybrid et al., 2021).Accuracy is often the most used metricrepresenting the comparison between predicted and actual labelseither true or false.(Gereme, Zhu, Ayall, & Alemu, 2021)(Rashid, Mustafa, & Saeed, 2018)The precision usedfor information retrieval. The precision score is calculated as out of thetotal positive results that are predicted by the model.precision shows the number of articles that are marked as true out of all the positively predicted (true) articles.(Khanam, Alwasel, Sirafi, & Rashid, 2021)The recall isTrue positive rate or True negative rate, it represents the number of articles predicted as true out of the total number of true articles. F1-score is acombination of precision and recall(B. Collins, D. T. Hoang, N. T. Nguyen, 2020).

4.Result and discussion

The results of our experiment on applying the five classifiers on the two sets of the corpuswere depicted in table 1 and table 2 respectively using the accuracy, precision, recall, and F1 score metrics to evaluate the classifiers.

Table 1. shows the performance evaluation of five classifiers with TF-IDFfeature on set 1 that includes fake news extracted directly from non-valid sources.

Table 1: Experimental results of recall, precision and F1-measure for 5 classifiers on 2 tests

Set 1				
Classifier	**Accuracy**	**Precision**	**Recall**	**F1**
Naïve Bayes	88.58	88.78	88.58	88.56
SVM	88.71	88.72	88.71	88.71
Decision Tree	80.44	80.54	80.44	80.42
Random Forest	86.34	86.34	86.33	86.33
Logistic Regression	87.58	87.61	87.57	87.57

Table 2. shows the performance evaluation of five classifiers with TF-IDF feature on set 2 that includes manipulated and modified news from real news to generate fake news

Table 2: Experimental results of recall, precision and F1-measure for 5 classifiers on set 2

Set 2				
Classifier	**Accuracy**	**Precision**	**Recall**	**F1**
Naïve Bayes	66.72	67.09	66.66	66.49
SVM	72.09	72.24	72.11	72.05
Decision Tree	82.19	82.23	82.18	82.18
Random Forest	83.26	83.27	83.26	83.26
Logistic Regression	78.89	79.08	78.92	78.86

Closer inspection of table 1 shows that the best result on the precisionmetrics was by NB with 88.78%, SVM outperformed the other classifiers in Recall, F1, and accuracy with 88.71%. however, in Table 2, the result indicates that Random Forest outstandingly performed comparing the other classifiers in set 2 with an accuracy 83.26%.

Interestingly, SVM accuracy from set 1 outperformed RF accuracy from set 2, these findings could be justified by the different nature of the two sets, set 2 includes machine-manipulated news-driven from real news and shares the same patterns.

Due to the absence ofsimilar work in Kurdish fake news detection, it was hard to benchmark our proposed system with similar work. By comparing our work with similar work in the Arabic language(Maysoon Alkhair, Karima Meftouh & Othman, 2019), we could notice that SVM scored the highest accuracy 95,35% against our work 88,71%, this slight difference it might be justified due to the nature of the language, the size of the dataset and the feature extraction used,more investigations and research should be done in this area.

5.Conclusion

In this paper, we discussed the first attempt to detect fake news in the Kurdish language, introducinga new Kurdish corpus and applied five classifiers.The outcomesindicated that the accuracy ofthe SVM classifieroutperformed the other classifiers with88,71% for set 1 and Random Forest scored the highest accuracy with 79.08% on a set 2 of data that has manipulated text. Fake news detection in a low-resourced language such as Kurdish is a new and critical issue for future research and there are still many unanswered questions and gaps.In the future, we look forwardto focusing on applyingensemble methods and sentiment analysisin detecting fake news in the Kurdish language.

References

1. Abdulrahman, R. O., Hassani, H., & Ahmadi, S. (2019). Developing a fine-grained corpus for a less-resourced language: The case of Kurdish. *ArXiv*, 106–109.
2. Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, *2020*. https://doi.org/10.1155/2020/8885861
3. Ahmadi, S. (2020). *KLPT – Kurdish Language Processing Toolkit*. 72–84. https://doi.org/10.18653/v1/2020.nlposs-1.11
4. Ahmed, H., Traore, I., & Saad, S. (2017). Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. *First International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, *10618*, 169–181. https://doi.org/10.1007/978-3-319-69155-8
5. AL-Saif, H., & Al-Dossari, H. (2018). Detecting and classifying crimes from arabic twitter posts using text mining techniques. *International Journal of Advanced Computer Science and Applications*, *9*(10), 377–387. https://doi.org/10.14569/IJACSA.2018.091046
6. Alanazi, S. S. (2020). *Arabic Fake News Detection In Social Media Using Readers ' Comments : Text Mining Techniques In Action*. *20*(9). https://doi.org/10.22937/IJCSNS.2020.20.09.4
7. Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I., & Gelbukh, A. (2020). "Bend the truth": Benchmark dataset for fake news detection in Urdu language and its evaluation. *Journal of Intelligent and Fuzzy Systems*, *39*(2), 2457–2469. https://doi.org/10.3233/JIFS-179905
8. B. Collins, D. T. Hoang, N. T. Nguyen, and D. H. (2020). 'Fake news types and detection models on social media: A state-of-the-art survey. In *Communications in Computer and Information Science*: *Vol.1178 CCIS* (pp. 562–573). Springer Nature Singapore. https://doi.org/https://doi.org/10.1007/978-981-15-3380-8_49
9. Elmurngi, E., & Gherbi, A. (2017). Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques. *DATA ANALYTICS 2017 : The Sixth International Conference on Data Analytics Detecting*, (c), 65–72.
10. Gereme, F., Zhu, W., Ayall, T., & Alemu, D. (2021). Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information (Switzerland)*, *12*(1), 1–9. https://doi.org/10.3390/info12010020
11. Hussain, M. G., Rashidul Hasan, M., Rahman, M., Protim, J., & Al Hasan, S. (2020). Detection of Bangla Fake News using MNB and SVM Classifier. *Proceedings - 2020 International Conference on Computing, Electronics and Communications Engineering, ICCECE 2020*, 81–85. https://doi.org/10.1109/iCCECE49321.2020.9231167
12. Hybrid, U., Hawks, H., & Feature, B. (2021). *Intelligent Detection of False Information in Arabic Tweets Machine Learning Models*.
13. Khalifa, M., & Hussein, N. (2019). Ensemble learning for irony detection in Arabic tweets. *CEUR Workshop Proceedings*, *2517*, 433–438.
14. Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake News Detection Using Machine Learning Approaches. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012040. https://doi.org/10.1088/1757-899x/1099/1/012040
15. Maakoul, O., Boucht, S., El Hachimi, K., & Azzouzi, S. (2020). Towards Evaluating the COVID'19 related Fake News Problem: Case of Morocco. *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science, ICECOCS 2020*. https://doi.org/10.1109/ICECOCS50124.2020.9314517
16. Maysoon Alkhair, Karima Meftouh, K. S., & Othman, and N. (2019). An Arabic Corpus of Fake News: Collection, Analysis and Classification. *7th International Conference on Arabic Language Processing: From Theory to Practice (ICALP 2019)*, (October), 292–302. https://doi.org/10.1007/978-3-030-32959-4
17. Mustafa, A. M., & Rashid, T. A. (2018). Kurdish stemmer pre-processing steps for improving information retrieval. *Journal of Information Science*, *44*(1), 15–27. https://doi.org/10.1177/0165551516683617
18. Nagoudi, E. M. B., Elmadany, A. R., Abdul-Mageed, M., Alhindi, T., & Cavusoglu, H. (2020). Machine generation and detection of arabic manipulated and fake news. *ArXiv*, 1–15.

19. Rashid, T. A., Mustafa, A. M., & Saeed, A. M. (2018). Automatic kurdish text classification using KDC 4007 dataset. *Lecture Notes on Data Engineering and Communications Technologies*, *6*, 187–198. https://doi.org/10.1007/978-3-319-59463-7_19
20. Sharma, N., Litoriya, R., Pratap Singh, H., & Sharma, D. (2021). *Modern Approach for the Significance Role of Decision Support System in Solid Waste Management System (SWMS)*. https://doi.org/10.1007/978-981-15-4936-6_67
21. Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004
22. Zhou, S., Cai, S., Zeng, C., & Wang, Z. (2020). A Novel Approach for Selecting Hybrid Features from Online News Textual Metadata for Fake News Detection. In *Lecture Notes in Networks and Systems* (Vol. 96). https://doi.org/10.1007/978-3-030-33509-0_14